PROFIS: Design of structurally-novel drug candidates by probing molecular fingerprint space with RNNs

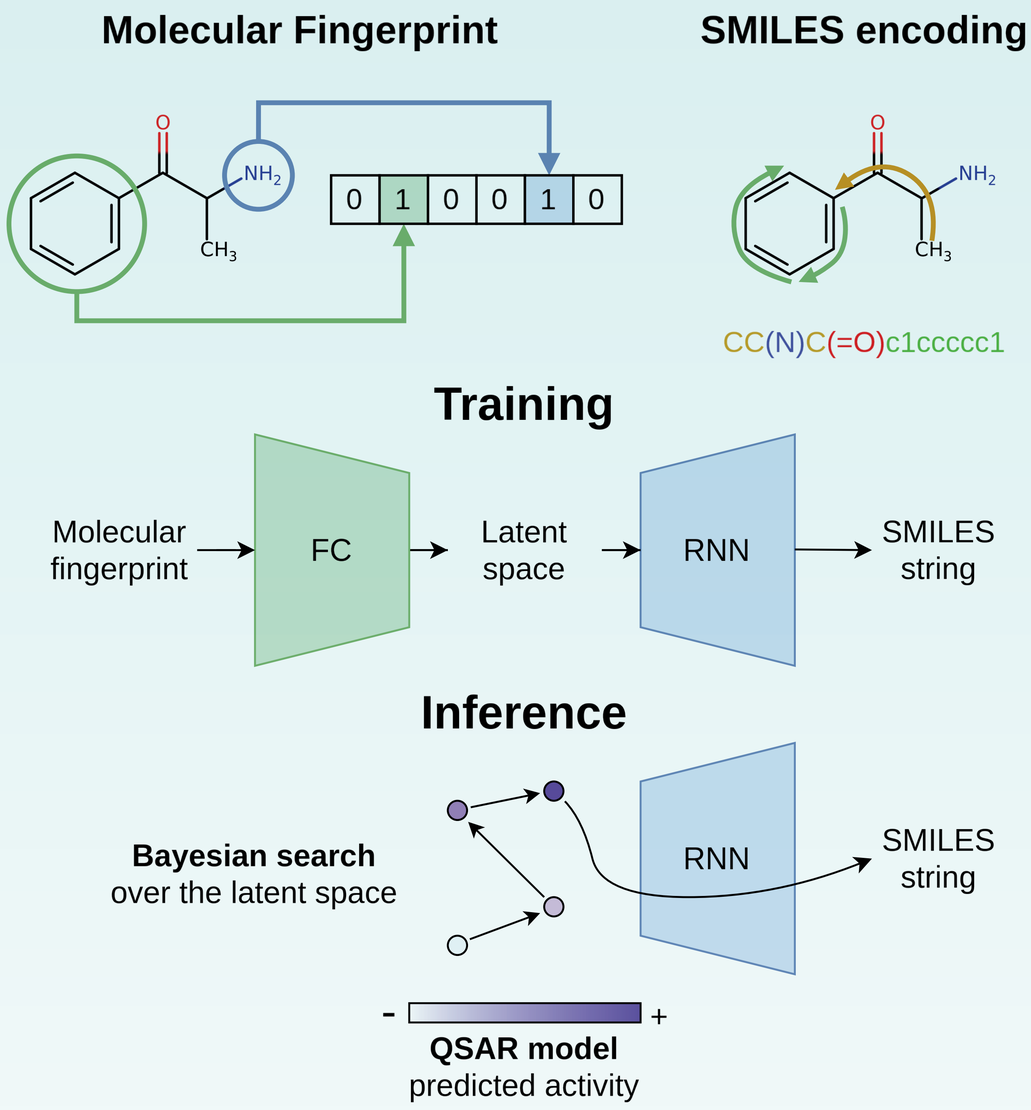
Hubert Rybka[†,‡] **Tomasz Danel**[‡] **Sabina Podlewska**[¶]

[†]Doctoral School of Exact and Natural Sciences, Jagiellonian University, Łojasiewicza 11, 30-348, Kraków, Poland
[‡]Faculty of Chemistry, Jagiellonian University, Gronostajowa 2, 30-387 Kraków, Poland
[¶]Maj Institute of Pharmacology, Polish Academy of Sciences, Smętna 12, 31-343 Kraków, Poland

ABSTRACT

PROFIS is a variational autoencoder capable of the design of structurally novel and target-focused compound libraries. The model relies on an RNN trained to decode embedded molecular fingerprints into SMILES strings. On this low-dimensional fingerprint embedding space, a biological activity predictor is then trained, enabling the identification of high-activity subspaces for a given drug target. In conjunction with this predictive model, a Bayesian search algorithm is used to identify the fingerprint embedding vectors which are expected to yield active structures upon decoding. **This study demonstrates the application of this protocol to generate candidate ligands of dopamine D2 receptor**. It emphasizes the effectiveness of our approach in scaffold hopping, which is is a method of designing ligands outside the already explored chemical space. **The PROFIS protocol is versatile and it can be employed for any biological target, given the availability of a dataset of known binders.**

METHODS

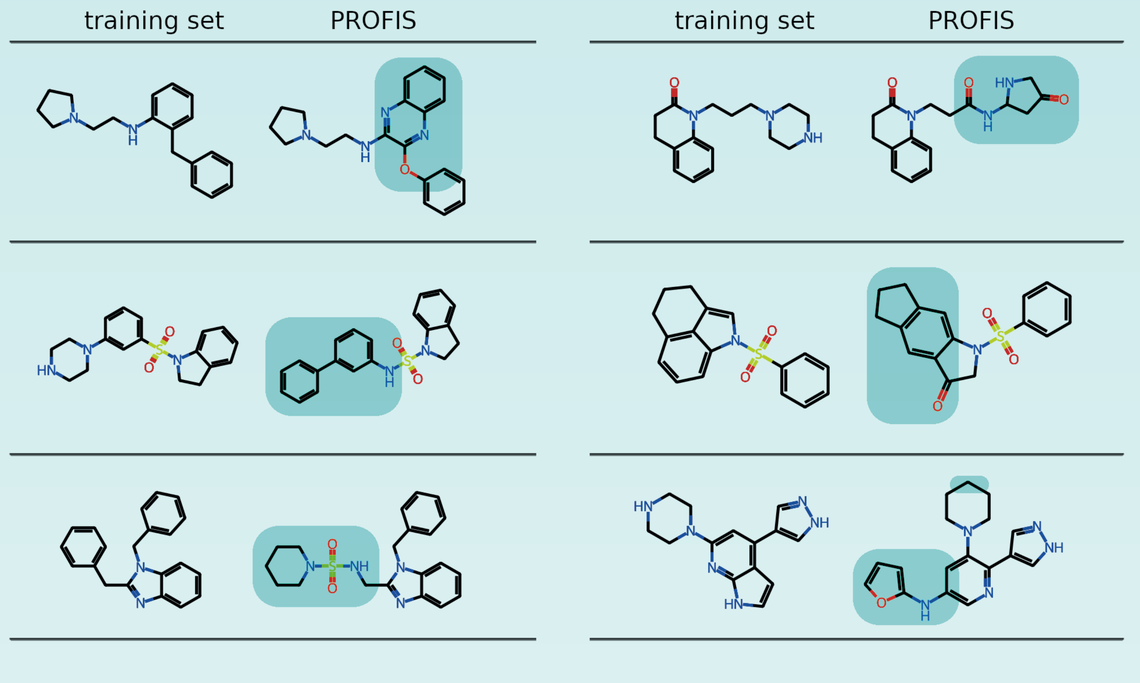


RESULTS

The cornerstone of our generative model is a low-dimensional embedding space of molecular FPs, which can be effectively searched for subspaces that represent structures of potent ligands. To enable this task, a QSAR model in the form of an activity classifier is trained on the embedded fingerprints.

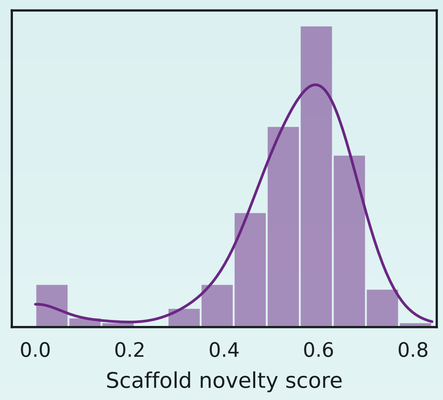
For this study, we sourced a dataset of 9391 D2R ligands with known Ki values from ChEMBL32 database and assigned those of Ki < 100 nM to the positive class, and the rest to the negative class. We developed a series of QSAR latent space classifiers and compared their performance with analogous models trained on uncompressed fingerprints. The latent classifiers (SVM, XGB and MLPs) succesfully aquired the capacity of identifying embeddings that corespond to D2R ligands of high biological activity.

		ECFP4		KRFP	
	model	ROC AUC	accuracy	ROC AUC	accuracy
latent	SVM	0.887 ± 0.009	**0.836** ± 0.007	0.885 ± 0.010	0.828 ± 0.008
	RF	0.770 ± 0.016	0.711 ± 0.009	0.777 ± 0.016	0.719 ± 0.008
	XGB	**0.891** ± 0.010	0.832 ± 0.015	0.891 ± 0.015	0.828 ± 0.008
	MLP	0.889 ± 0.010	0.825 ± 0.013	**0.893** ± 0.012	**0.830** ± 0.010
baseline	SVM	0.906 ± 0.008	**0.853** ± 0.010	0.905 ± 0.011	0.844 ± 0.010
	RF	0.797 ± 0.017	0.724 ± 0.010	0.822 ± 0.017	0.732 ± 0.010
	XGB	**0.910** ± 0.009	0.838 ± 0.010	**0.924** ± 0.010	**0.855** ± 0.012
	MLP	0.894 ± 0.011	0.834 ± 0.009	0.911 ± 0.010	0.842 ± 0.020

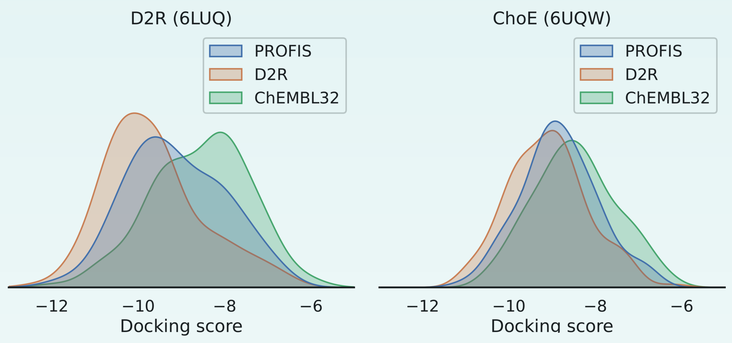


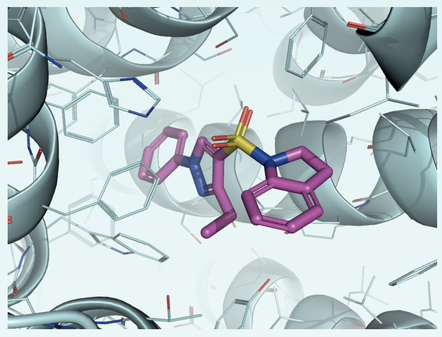
Generated compounds

- The structures produced by PROFIS exemplify the outcome of a classical scaffold-hopping strategy applied to molecules from the D2R ligands set.
- Majority of the molecules generated by PROFIS (~90%) are based on novel scaffolds, with no similar scaffolds (Tanimoto distance < 0.4) found in the training set.



Docking study





- The candidate D2R ligands generated by PROFIS exhibit stronger D2R binding affinities than a random CHEMBL32 database sample (Mann–Whitney U test, $p < 0.00005$).
- If docked to a structurally-unrelated protein (ChoE, bacterial acetylcholinesterase), docking score distributions of the random database sample, PROFIS-generated molecules and known D2R ligands overlap. This serves as a proof of our model's capacitiy to generate target-focused compound libraries.

Contact us

https://github.com/hubertrybka/profis

hubert.rybka@doctoral.uj.edu.pl
tomasz.danel@uj.edu.pl
smusz@if-pan.krakow.pl

code